МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ
(Вычислительные системы)
1978 год Выпуск 74

УДК 681.3.06:621.391

ОБ ОДНОМ ПОДХОДЕ К ПРОГНОЗИРОВАНИЮ ПЕРСПЕКТИВНОСТИ ОБЪЕКТОВ

А.Н.Манохин

§1. Постановка задачи

Пусть дана некоторая совокупность объектов, каждый из которых принадлежит к одному из двух образов. Под перспективностью понимается вероятность данного объекта попасть в фиксированный (будем считать, в первый) образ. В рамки формальной постановки задачи, рассматриваемой в работе, укладывается ряд содержательных задач экономики, планирования разведочных работ, планирования вложения средств, организации некоторых процессов обслуживания. Решение таких задач имеет большое прикладное значение.

В первой части этой работы описывается содержательная постановка задачи и предлагается формальная статистическая модель. Во второй части в рамках предложенной модели исследуются вопросы существования оптимальной стратегии для случая известных распределений. В третьей части на основе подхода, предложенного в [1], упорядочение строится по обучающей выборке и предлагаются алгоритмы, реализующие этот подход. Рассмотрим вначале некоторые содержательные иллюстрации тех постановок, которые описаны в работе.

ПРИМЕР 1. Исследуется пожароопасность в шахтах некоторого бассейна на основании известных характеристик $x_1, \ldots, x_n$. Эти характеристики описывают технологические, экономические, природные условия добычи. Известны данные о возникновении пожаров за некоторый период. На противопожарные мероприятия отпускается некоторая сумма C. Необходимо распределить эти средства по шахтам. Комплекс противопожарных мероприятий для шахт имеет стоимость C_1; пусть ν — общее количество шахт. Если $C/C_1 \ll \nu$, то необходимо принять решение о том, какие из ν шахт должны быть обеспечены этим комплексом. Причем желательно вложить средства наиболее оптимальным образом, т.е. предотвратить наибольшее число пожаров.

ПРИМЕР 2. Производятся геологоразведочные работы на участках некоторого района. Каждый из этих участков описывается характеристиками (геологическими, геохимическими и геофизическими) $x_1,\dots,x_n$. Определена целевая характеристика x_0, которая принимает два значения: 0 - отсутствие полезного ископаемого, 1 - наличие его. Определение этой характеристики, как правило, требует больших затрат C_1 (например, это могут быть буровые работы). Если средства на определение характеристики x_0 ограничены суммой C и $C/C_1 \ll \nu$ (ν - количество участков района), то необходимо опять же организовать вложение средств наиболее оптимальным образом, т.е. на основании признаков $x_1,\dots,x_n$ мы должны выбрать μ участков ($\mu=[C/C_1]$) таким образом, чтобы среди них оказалось как можно больше содержащих ископаемое.

ПРИМЕР 3. Проводится медицинское обследование большого коллектива людей с целью ранней диагностики заболевания.

Обследование делится на два этапа. Первый - сбор простейших характеристик $x_1,\dots,x_n$ (например, результаты осмотра одним врачом, анкетные данные, условия труда, социологические данные и т.д.). Такое обследование может быть проведено на всей совокупности. Второй этап - определение целевой характеристики x_0: имеется ли ранняя стадия заболевания или нет. Для точного определения x_0 необходимо провести сложное и дорогостоящее обследование (это проведение сложных лабораторных анализов, клиническое обследование). Подобное обследование нельзя провести для каждого члена совокупности как из-за ограниченности средств (мест в поликлинике, мощности лабораторий), так и вследствие больших потерь рабочего времени, вызываемых таким обследованием. Тогда встает задача выбора μ объектов из ν наиболее оптимальным образом, а именно: из общей совокупности людей необходимо выбрать максимальное количество действительно больных, подлежащих клиническому обследованию.

Отметим общие моменты в постановке вышеизложенных задач:

1) предполагается наличие обучающей выборки, на основании которой мы получаем информацию о связях признаков с целевым признаком;

2) требуется принимать решение о наиболее выгодном распределении некоторых ограниченных ресурсов;

3) допускается, что признаки могут быть замерены в различных шкалах;

4) предполагается, что объем обучающей выборки может быть незначителен по сравнению с размерностью пространства и фиксирован, т.е. количество наблюдений не может произвольно увеличиваться.

Предлагаемая далее формальная постановка задачи будет служить математической моделью для изложенных выше содержательных задач.

Пусть $\bar{x}$ – вектор пространства $X_1 * X_2 * \ldots * X_n$ и пусть существуют условные распределения $P(\bar{x}/1)$ и $P(\bar{x}/0)$ (последние могут быть и неизвестны). Каждый объект a из генеральной совокупности описывается набором значений признаков $X_1(a), \ldots, X_n(a)$, которые определяют вектор $\bar{x}$. Совокупность, подлежащая упорядочению, содержит ν объектов $\{a_1, \ldots, a_\nu\}$, которым сопоставляется совокупность векторов в признаковом пространстве $\{\bar{x}_1, \ldots, \bar{x}_\nu\}$, где $\bar{x}_i = \bar{X}(a_i) = (X_1(a), \ldots, X_n(a))$. Чтобы отличать функцию от конкретного значения функции, будем обозначать вектор признаков через $\bar{x}$, когда объект-аргумент a отсутствует, и через $\bar{X}(a)$ – когда он присутствует. Решающее правило F должно из заданных ν объектов выбрать μ, где $\mu<\nu$ и константа μ определяется экономическими ограничениями, примеры которых были приведены. Выбор осуществляется на основе вектор-признаков.

ОПРЕДЕЛЕНИЕ I. Отображение $F:(\bar{x}_1, \ldots \bar{x}_\nu) \to (i_1, \ldots, i_\mu)$ назовем решающим правилом F.

Таким образом, F задает набор номеров $\{i_1, \ldots, i_\mu\}$, которому соответствует набор объектов $\{a_{i_1}, \ldots, a_{i_\mu}\}$. Эти объекты в соответствии с правилом F объявляются наиболее перспективными.

Качеством конкретного выбора $a_{i_1}, \ldots, a_{i_\mu}$ будем считать величину λ – количество объектов первого образа в совокупности $\{a_{i_1}, \ldots, a_{i_\mu}\}$, $\lambda \le \mu < \nu$.

Пусть $X_0(a_i) = 1$, если a_i принадлежит первому образу и $X_0(a_i) = 0$, если a_i принадлежит второму образу. Тогда $\lambda = \sum_{k=1}^{\mu} X_0(a_{i_k})$. Вектор $\bar{X}(a_i)$ сопоставляется объекту a_i в соответствии с распределением $P(\bar{x}/1)$, если $X_0(a_i)=1$, и $P(\bar{x}/0)$, если $X_0(a_i) = 0$. Тогда решающее правило определяет случайную величину λ полностью, если дополнительно определены априорные вероятности $P(1) = P(X_0(a_i) = 1)$ и $P(0) = P(X_0(a_i) = 0)$. Можно рассматривать случаи, когда механизм задания значения $X_0(a_i)$ не является случайным, т.е. нельзя говорить об априорных вероятностях. Однако результаты, приводимые ниже, остаются справедливыми и для этого случая. Тогда мы считаем, что значения $X_0(a_i)$, $a_i \in \{a_1, \dots, a_\nu\}$, заданы некоторым образом, значения $\bar{X}(a_i)$ определяются, как и раньше, в соответствии с $P(\bar{x}/1)$ и $P(\bar{x}/0)$, и решающее правило F по-прежнему определяет случайную величину λ.

Критерием качества решающего правила F будем считать среднее число объектов первого образа в выделенной совокупности $M\lambda(F,\bar{P})$, $\bar{P} = \{P(1), P(0), P(\bar{x}/1), P(\bar{x}/0)\}$.

Далее будет исследован вопрос о выборе оптимального правила, когда распределения $P(\bar{x}/1)$ и $P(\bar{x}/0)$ известны, и рассмотрена возможность построения правила по обучающей выборке.

Приведем теперь вторую формальную постановку задачи. Пусть так же, как и в первой постановке, для каждого объекта a_i определены значения характеристики $X_0(a_i)$. На множестве объектов $\{a_1, \dots, a_\nu\}$ необходимо установить порядок $a_{i_1} \prec a_{i_2} \prec a_{i_3} \prec \dots \prec a_{i_\nu}$. Константу μ, определяющую критерий качества, теперь полагаем незаданной. Необходимо в целом определить качество упорядочения. Желательно, чтобы в ряду $X_0(a_{i_1}), \dots, X_0(a_{i_\nu})$ единицы концентрировались в начале, а нули в конце ряда, и тогда лучший порядок определится следующим рядом значений для $X_0(a_i)$: $\underbrace{1, \dots, 1}_{q \text{ раз}}$, $\underbrace{0, \dots, 0}_{\nu-q \text{ раз}}$, а худший – $\underbrace{0, \dots, 0}_{\nu-q \text{ раз}}$, $\underbrace{1, \dots, 1}_{q \text{ раз}}$.

Определим критерий качества для любого упорядочения. Пусть

$$\Delta(a_i) = \begin{cases} 0, \text{ если } X_0(a_i) = 0, \\ \pi(a_i), \text{ если } X_0(a_i) = 1 \end{cases},$$

где $\pi(a_i)$ – номер места в упорядочении, на которое поставлен объект a_i. В качестве критерия качества упорядочения предлагается $\varphi(\Pi) = \sum_{i=1}^{\nu} \Delta(a_i)$, где Π – набор $(i_1, \dots, i_\nu)$, определяющий упорядочение. Упорядочение тем лучше, чем меньше $\varphi(\Pi)$.

Рассмотрим содержательную трактовку этого критерия. Пусть за единицу времени может быть обработан один объект, например, пробурена скважина. Как только мы установим принадлежность этого объекта к первому образу, то этот объект начинает приносить доход τ в единицу времени. Если же объект принадлежит ко второму образу, то он не приносит дохода и средства, затраченные на его разработку, использованы впустую. За время T $(T>\nu)$ объект принесет доход $[T+1-\pi(a_i)]\tau$ и минимальный общий доход будет равен

$$[(T-(\nu-q))+(T-(\nu-q+1))+\ \dots\ +(T-(\nu-1))]\tau = (T-\nu+\frac{q+1}{2})q\tau ,$$

а максимальный доход –

$$[T+(T-1)+\dots+(T-(q-1))]\tau = (T-\frac{q-1}{2})q\tau .$$

Потери в доходе есть разница между максимальным доходом и полученным; они равны

$$(T-\frac{q-1}{2})q\tau-(qT+q-\sum_{i=1}^{\nu}\Delta(a_i))\tau = \left[\varphi(\Pi)-\frac{(q+1)q}{2}\right]\tau .$$

Легко видеть, что для потерь выполняется неравенство

$$0 \le \left[\varphi(\Pi)-\frac{(q+1)q}{2}\right]\tau \le (\nu-q)q\tau .$$

Мы видим, что потери линейно связаны с величиной $\varphi(\Pi)$. Можно рассмотреть нормированный критерий

$$\tilde{\varphi}(\Pi) = \frac{\varphi(\Pi)-\frac{(q+1)q}{2}}{(\nu-q)q} ,$$

который представляет отношение полученных потерь к максимально возможным потерям.

ОПРЕДЕЛЕНИЕ 2. Критерием качества решающего правила F, которое набору $\bar{x}_1,\dots,\bar{x}_\nu$ $(\bar{x}_i=\bar{X}(a_i))$ сопоставляет набор $(i_1, i_2,\dots,i_\nu)$, будем считать среднее значение $\varphi(\Pi)$, т.е. $\varphi(F,\bar{P}) = M\varphi(\Pi)$.

§2. Оптимальное решающее правило при известных распределениях

Рассмотрим задачу нахождения оптимального решающего правила в случае, когда $P(\bar{x}/1)$ и $P(\bar{x}/0)$ известны. Будет найдено решающее правило F_0 такое, что

$$M\lambda(F_0,\bar{P}) \ge M\lambda(F,\bar{P})$$

для каждого F при каждой фиксированной стратегии природы

$$\bar{P} = \{P(\bar{x}/1),\ P(\bar{x}/0),\ P(1),\ P(0)\}.$$

Результаты, полученные в этом параграфе, представляют аналог леммы Неймана-Пирсона для рассматриваемой задачи [2,3].

Определим функцию правдоподобия $g(\bar{x})$ для объекта a такого, что $\bar{x}(a) = \bar{x}$, следующим образом:

$$\frac{f(\bar{x}/1)}{f(\bar{x}/0)},$$

если распределения $P(\bar{x}/1)$ и $P(\bar{x}/0)$ имеют плотности $f(\bar{x}/1)$ и $f(\bar{x}/0)$;

$$\frac{p(\bar{x}/1)}{p(\bar{x}/0)},$$

если распределения $P(\bar{x}/1)$ и $P(\bar{x}/0)$ дискретны (здесь $\bar{x}$ – один из конечного либо счетного множества исходов);

$$\frac{p(\bar{x}_Д/1)\cdot f(\bar{x}_Н/\bar{x}_Д,1)}{p(\bar{x}_Д/0)\cdot f(\bar{x}_Н/\bar{x}_Д,0)},$$

если $P(\bar{x}/1)$ и $P(\bar{x}/0)$ непрерывно-дискретны.

Непрерывно-дискретный вектор состоит из двух компонент: дискретной – $\bar{x}_Д$ и непрерывной – $\bar{x}_Н$. Для $\bar{x}_Д$ определены вероятности $p(\bar{x}_Д/1)$, $p(\bar{x}_Д/0)$, а для $\bar{x}_Н$ определены условные плотности $f(\bar{x}_Н/\bar{x}_Д,1)$, $f(\bar{x}_Н/\bar{x}_Д,0)$.

Решающее правило F_0 определим следующим образом: объекты a_1, $a_2,\ldots,a_\nu$ упорядочиваются по убыванию функции $g(a_i)$, и первые μ объектов включаются в искомую совокупность.

ТЕОРЕМА I. Для каждого решающего правила F и любой стратегии природы $\bar{P}$ справедливо неравенство $M\lambda(F_0,\bar{P}) \geq M\lambda(F,\bar{P})$.

ДОКАЗАТЕЛЬСТВО. Рассмотрим случай, когда заданы априорные вероятности $P(1)$ и $P(0)$. Пусть распределения $P(\bar{x}/1)$ и $P(\bar{x}/0)$ имеют плотности $f(\bar{x}/1)$ и $f(\bar{x}/0)$ соответственно. Решающее правило F каждому набору векторов признаков $\{\bar{x}_1,\ldots,\bar{x}_\nu\}$ сопоставляет μ номеров $\{i_1,\ldots,i_\mu\}$, $1 \leq i_j < \nu$, т.е.

$$F(\bar{x}_1,\ldots,\bar{x}_\nu) = (i_1,\ldots,i_\mu).$$

Пусть $X^{\nu\circ a} = X^a \circ X^a \circ \ldots \circ X^a$ – прямое произведение признаковых пространств (пространство векторов признаков); T – множество всех различных векторов $\bar{\omega}$ из 0 и I длины ν (единица на i-м месте в векторе $\bar{\omega}$ означает, что $X_0(a_i) = 1$, а 0 – что $X_0(a_i) = 0$); ω_i есть i-я координата вектора $\bar{\omega}$.

Тогда

$$M\lambda(F,\bar{P}) = \int\limits_{X^{\nu \cdot m}}\!\!\ldots\!\int \sum_{\bar{\omega}\in T}\left[\prod_{i=1}^{\nu} P(\omega_i)\, f(\bar{x}_i/\omega_i) \times \right.$$

$$\left. \times \left(\sum_{j\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \omega_j\right)\right] d(\bar{x}_1,\ldots,\bar{x}_\nu).$$

При $f(\bar{x}) = P(1)\cdot f(\bar{x}/1) + P(0)\cdot f(\bar{x}/0)$ имеем

$$M\lambda(F,\bar{P}) = \int\limits_{X^{\nu \cdot m}}\!\!\ldots\!\int \sum_{\bar{\omega}\in T}\left[\prod_{i=1}^{\nu} f(\bar{x}_i)\cdot P(\omega_i/\bar{x}_i)\right) \times$$

$$\times \left(\sum_{j\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \omega_j\right)\Bigg] d(\bar{x}_1,\ldots,\bar{x}_\nu) = \int\limits_{X^{\nu \cdot m}}\!\!\ldots\!\int \prod_{i=1}^{\nu} f(\bar{x}_i) \times$$

$$\times \left[\left(\sum_{\bar{\omega}\in T} \left(\prod_{i=1}^{\nu} P(\omega_i/\bar{x}_i)\right)\right)\left(\sum_{j\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \omega_j\right)\right] d(\bar{x}_1,\ldots,\bar{x}_\nu).$$

Рассмотрим выражение

$$\sum_{\bar{\omega}\in T}\left(\prod_{i=1}^{\nu} P(\omega_i/\bar{x}_i)\right)\cdot\left(\sum_{j\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \omega_j\right).$$

При фиксированном бинарном векторе $\bar{\omega}$ выражение $\prod_{i=1}^{\nu} P(\omega_i/\bar{x}_i)$ определяет вероятность этого вектора в следующей схеме: в i-м испытании с вероятностью $P(1/\bar{x}_i)$ имеет место успех и с вероятностью $P(0/x_i)$ – неуспех, испытания независимы, $\sum_{j\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \omega_j$ – случайная величина (обозначим ее через ξ), которая дает количество успехов в выделенном множестве испытаний. Тогда $\xi = \sum_{i\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \xi_i$, где $\xi_i = 1$ – в случае успеха и $\xi_i = 0$ – в случае неуспеха в i-м испытании. Следовательно,

$$\sum_{\bar{\omega}\in T}\left(\prod_{i=1}^{\nu} P(\omega_i/\bar{x}_i)\right)\cdot\left(\sum_{j\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} \omega_j\right) = M\xi = \sum_{i\in F(\bar{x}_1,\ldots,\bar{x}_\nu)} M\xi_i .$$

Но $M\xi_i = 1\cdot P(1/\bar{x}_i) + 0\cdot P(0/\bar{x}_i) = P(1/\bar{x}_i)$, и подынтегральное выражение равно

$$\sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu}(P(\omega_i/\bar x_i))\right)\left(\sum_{j\in F(\bar x_1,\dots,\bar x_\nu)}\omega_j\right) =$$

$$= \sum_{j\in F(\bar x_1,\dots,\bar x_\nu)} P(1/\bar x_j) \ . \qquad (I)$$

Таким образом, выбор решающего правила F при фиксированном наборе векторов признаков $\{\bar x_1,\dots,\bar x_\nu\}$ определяет множество индексов, по которому ведется суммирование в (I). Далее, $P(1/\bar x)\geq P(1/\bar y)$ тогда и только тогда, когда $g(\bar x)\geq g(\bar y)$. Это следует из того, что

$$P(1/\bar x) = \frac{P(1)\cdot f(\bar x/1)}{P(1)\cdot f(\bar x/1) + P(0)\cdot f(\bar x/0)} = \frac{1}{1+\dfrac{P(0)\cdot f(\bar x/0)}{P(1)\cdot f(\bar x/1)}} \ .$$

Правило F_0 определяет область суммирования следующим образом. Числа $P(1/\bar x_j)$ упорядочиваются по убыванию, и в область $F(\bar x_1,\dots,\bar x_\nu)$ включаются первые μ индексов. Тогда для каждого правила F и каждого набора векторов признаков $\{\bar x_1,\dots,\bar x_\nu\}$ выполняется неравенство:

$$\sum_{j\in F(x_1,\dots,x_\nu)} P(1/\bar x_j) \geq \sum_{j\in F(\bar x_1,\dots,\bar x_\nu)} P(1/x_j) \ ,$$

следовательно, для подынтегрального выражения справедливо

$$\sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu}P(\omega_i/\bar x_i)\right)\cdot\left(\sum_{j\in F_0(\bar x_1,\dots,\bar x_\nu)}\omega_j\right) \geq$$

$$\geq \sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu}P(\omega_i/\bar x_i)\right)\cdot\left(\sum_{j\in F(\bar x_1,\dots,\bar x_\nu)}\omega_j\right).$$

Используя простое свойство интегрирования (если $f_1(\bar x)\geq f_2(\bar x)$ для каждого $\bar x$, то $\int\limits_D\dots\int f_1(\bar x)d\bar x \geq \int\limits_D\dots\int f_2(\bar x)d\bar x$), получаем $M\lambda(F_0,\bar P)\geq M\lambda(F,\bar P)$ для каждого F и $\bar P$. Доказательство закончено.

Рассмотрим теперь вопрос об оптимальности правила F_0 , определенного выше, в случае, когда природа определения принадлежности объекта к одному из классов не является случайной, т.е. отсутствуют априорные вероятности $P(0)$ и $P(I)$.

Итак, пусть нам известно, что среди объектов $a_1,\dots,a_\nu$ q принадлежат к первому образу и $\nu-q$ – ко второму. Нам неизвестны номера объектов, принадлежащих к первому образу, зато известно, что для каждого объекта a_i, принадлежащего к первому образу, реализуется случайная векторная величина с законом распределения $P(\bar{x}/1)$ и значение $\bar{x}_i$ нам сообщается, а для объектов второго образа реализуется случайная величина $\bar{x}$ с распределением $P(\bar{x}/0)$. В этом случае стратегия природы есть набор $\bar{P} = \{ P(\bar{x}/1), P(\bar{x}, 0), i_1, \dots, i_q \}$, где $i_1, \dots, i_q$ – номера объектов, принадлежащих к первому образу. Очевидно, что решающее правило F_0 в этой постановке уже не будет обладать таким сильным свойством, как $M\lambda(F_0, \bar{P}) \geq M\lambda(F, \bar{P})$ для каждого решающего правила F и стратегии природы $\bar{P}$.

Действительно, пусть

$$\int_D f_1(\bar{x})d\bar{x} > 0, \text{ где } \quad D = \{\bar{x} | 0 < \frac{f_1(\bar{x})}{f_2(\bar{x})} < \infty\},$$

и F_1 – решающее правило, которое любому набору векторов признаков сопоставляет набор номеров $\{i'_1, \dots, i'_\mu\}$ следующим образом: при $\mu < q$ сопоставляется μ первых номеров из набора $\{i_1, \dots, i_q\}$, при $\mu \geq q$ сопоставляется весь набор $\{i_1, \dots, i_q\}$ и произвольно остальные $\mu - q$ векторов.

Очевидно, что в случае $\mu \leq q$ имеет место $M\lambda(F_1, \bar{P}) = \mu$, а в случае $\mu \geq q$ справедливо $M\lambda(F_1, \bar{P}) = q$ и $M\lambda(F_1, \bar{P}) \geq M\lambda(F_0, \bar{P})$.

Но в то же время ясно, что решающее правило F_1 не может представлять практического интереса, так как оно определяется через набор $\{i_1, \dots, i_q\}$, который нам неизвестен.

Введем теперь определение стратегии, обеспечивающей минимакс потерь ([I, гл.VI, §3]).

ОПРЕДЕЛЕНИЕ 3. Решающее правило $\tilde{F}$ обеспечивает минимакс потерь, если для каждой стратегии природы $\bar{P}$ и для каждого решающего правила F из некоторого класса решающих правил выполняется неравенство

$$\sup_{\bar{P}} (\sup_{F_1} M\lambda(F_1, \bar{P}) - M\lambda(F, \bar{P})) \geq \sup_{\bar{P}} (\sup_{F_1} M\lambda(F_1, \bar{P}) - M\lambda(\tilde{F}, \bar{P})).$$

ТЕОРЕМА 2. Решающее правило, сформулированное выше на основании упорядочения по функции правдоподобия $g(\bar{x})$, обеспечивает минимакс потерь.

ДОКАЗАТЕЛЬСТВО этой теоремы аналогично доказательству теоремы I и поэтому опускается.

Теперь перейдем к формулировке оптимальной стратегии и ее свойствам во второй постановке задачи. Предварительно будет доказана

ЛЕММА. Пусть ξ_i - случайная величина, которая принимает значение l_i с вероятностью p_i и 0 с вероятностью $1-p_i$, $l_i \in \{1,\dots,\nu\}$, $l_i \neq l_j$ при $i \neq j$. Набор $\{p_1,\dots,p_\nu\}$ фиксирован. Тогда минимум $M\xi = M(\xi_1+\dots+\xi_\nu)$ достигается, если $l_i < l_j$ при $p_i > p_j$.

ДОКАЗАТЕЛЬСТВО. Пусть оптимальное задание набора $\{l_1,\dots,l_\nu\}$ не удовлетворяет условию. Тогда существует пара i, j такая, что $p_i > p_j$ и $l_i > l_j$. Рассмотрим новый набор $\{l'_1,\dots,l'_\nu\}$, который задан следующим образом: $l'_i = l_j$, $l'_j = l_i$ и $l'_k = l_k$, если $k \neq i,j$.

Пусть $\xi'_1,\dots,\xi'_\nu$ - соответствующие этому набору случайные величины. Тогда

$$M\xi - M\xi' = M(\xi_i + \xi_j) - M(\xi'_i + \xi'_j) = p_i l_i + p_j l_j - p_i l_j - p_j l_i =$$

$$= p_i(l_i - l_j) + p_j(l_j - l_i) = (l_i - l_j)(p_i - p_j) > 0,$$

т.е. $M\xi > M\xi'$, что противоречит предположению об оптимальности набора $\{l_1,\dots,l_\nu\}$.

Перейдем к доказательству теоремы об оптимальности решающего правила F_0, задаваемого функцией правдоподобия. Правило F_0 сопоставляет набору векторов признаков $\{\bar{x}_1,\dots,\bar{x}_\nu\}$ упорядоченный набор $\{i_1,\dots,i_\nu\}$, $i_j \in \{1,\dots,\nu\}$, который задается следующим образом. Если $g(\bar{x}_i) > g(\bar{x}_j)$, то i в наборе предшествует j; если $g(\bar{x}_i) = g(\bar{x}_j)$, то вероятности событий "i предшествует j" и "j предшествует i" одинаковы и равны $\frac{1}{2}$.

ТЕОРЕМА 3. Решающее правило F_0 является оптимальным, т.е. для каждого F и каждой стратегии природы $\bar{P}$ выполняется неравенство $M\varphi(F_0,\bar{P}) \leq M\varphi(F,\bar{P})$.

ДОКАЗАТЕЛЬСТВО. Рассмотрим случай, когда определены априорные вероятности $P(0)$, $P(I)$ и распределения имеют плотности $f(\bar{x}/0)$, $f(\bar{x}/1)$. Тогда $M\varphi(\Pi) = \sum_{i=1}^{\nu} M\Delta(a_i)$, где $\varphi(\Pi)$ и $\Delta(a_i)$ были определены выше. Следовательно,

$M_\varphi(\Pi)=$

$$=\int_{X^{\nu\cdot m}}\!\!\cdots\int \sum_{\bar\omega\in T}\left[\left(\prod_{i=1}^{\nu} P(\omega_i)f(\bar x_i/\omega_i)\right)\cdot\varphi\left(F(\bar x_1,\ldots,\bar x_\nu)\right)\right]d(\bar x_1,\ldots,\bar x_\nu)=$$

$$=\int_{X^{\nu\cdot m}}\!\!\cdots\int \prod_{i=1}^{\nu} f(\bar x_i)\left[\sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu} P(\omega_i/\bar x_i)\,\varphi\left(F(\bar x_1,\ldots,\bar x_\nu)\right)\right)\right]d(\bar x_1,\ldots,\bar x_\nu).$$

Рассмотрим далее выражение

$$\sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu} P(\omega_i/\bar x_i)\right)\varphi\left(F(\bar x_1,\ldots,\bar x_\nu)\right). \qquad (2)$$

Положим $p_i = P(1/\bar x_i)$ и заметим, что выбор порядка $\Pi = F(\bar x_1,\ldots,\bar x_\nu)$ означает задание набора $\{l_1,\ldots,l_\nu\}$, $l_i\in\{1,\ldots,\nu\}$, $l_i\neq l_j$ при $i\neq j$. Выражение (2) есть не что иное, как матожидание случайной величины ξ, определяемой в лемме по наборам $\{p_1,\ldots,p_\nu\}$ и $\{l_1,\ldots,l_\nu\}$. На основании леммы необходимо задать набор $\{l_1,\ldots,l_\nu\}$ таким образом, чтобы выполнялось условие: если $p_i > p_j$, то $l_i < l_j$.

Если мы зададим набор $\{i_1,\ldots,i_\nu\}$ в соответствии с функцией правдоподобия g, то требуемое условие будет выполнено. Действительно, если $P(1/\bar x_i) > P(0/\bar x_i)$, то

$$\frac{P(1)\cdot f(\bar x_i/1)}{P(1)\cdot f(\bar x_i/1)+P(0)\cdot f(\bar x_i/0)} > \frac{P(1)\cdot f(\bar x_j/1)}{P(1)f(\bar x_j/1)+P(0)f(\bar x_j/0)},$$

что эквивалентно

$$\frac{1}{1+\dfrac{P(0)f(\bar x_i/0)}{P(1)f(\bar x_i/1)}} > \frac{1}{1+\dfrac{P(0)f(\bar x_j/0)}{P(1)f(\bar x_j/1)}}$$

и

$$\frac{f(\bar x_i/1)}{f(\bar x_i/0)} > \frac{f(\bar x_j/1)}{f(\bar x_j/0)}.$$

Это условие, по определению F_0, означает, что i предшествует j в $\{i_1,\ldots,i_\nu\} = F(\bar x_1,\ldots,\bar x_\nu)$.

Таким образом, если упорядочение задавать в соответствии с решающим правилом F_0, то для каждого $\bar x_1,\ldots\bar x_\nu$ и каждого F вы -

полняется неравенство

$$\sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu} P(\omega_i/\bar x_i)\ \varphi(F_0(\bar x_1,\dots,\bar x_\nu))\right)\le$$

$$\le\sum_{\bar\omega\in T}\left(\prod_{i=1}^{\nu} P(\omega_i/\bar x_i)\ \varphi(F(\bar x_1,\dots,\bar x_\nu))\right).$$

Так как оно выполняется для каждого набора $x_1,\dots,x_\nu$, то

$$\int\limits_{X^{\nu\cdot m}}\!\!\dots\!\int \prod_{i=1}^{\nu} f(\bar x_i)\Big[\sum_{\bar\omega\in T}\Big(\prod_{i=1}^{\nu} P(\omega_i/\bar x_i)\times$$

$$\times\varphi(F_0(\bar x_1,\dots,\bar x_\nu))\Big)\Big]\,d(\bar x_1,\dots,\bar x_\nu)\le$$

$$\le\int\limits_{X^{\nu\cdot m}}\!\!\dots\!\int \prod_{i=1}^{\nu} f(\bar x_i)\Big[\sum_{\bar\omega\in T}\Big(\prod_{i=1}^{\nu} P(\omega_i/\bar x_i)\times$$

$$\times\varphi(F(\bar x_1,\dots,\bar x_\nu))\Big)\Big]d(\bar x_1,\dots,\bar x_\nu).$$

Доказательство закончено.

В случае, когда природа появления первого и второго образов не случайна, т.е. априорные вероятности не заданы, можно аналогично доказать оптимальность F_0 в смысле минимакса потерь.

§3. Упорядочение на основе обучающей выборки. Оценки равномерного уклонения

Будем рассматривать задачу установления порядка $\Pi=\{i_1,\ \ ,i_\nu\}$ на совокупности объектов $\{a_1,\dots,a_\nu\}$, когда распределения $P(\bar x/1)$, $P(\bar x/0)$ неизвестны, но известна обучающая выборка $\{b_1,\dots,b_n\}$. Предварительно нам потребуются некоторые вспомогательные свойства критерия $M\varphi(F,\bar P)$. Рассмотрим вторую постановку задачи, т.е. тот случай, когда порядок нужно установить на всем множестве $\{a_1,\dots,a_\nu\}$.

Пусть $\phi_1(\bar x)=\{\bar y|\bar y\succ\bar x\}$, $\phi_2(\bar x)=\{\bar y|\bar y\sim\bar x\}$, где значками $\succ$ и $\sim$ обозначены отношения предпочтения и эквивалентности. Вероятности множества A, определяемые распределениями $P(\bar x/1)$ и $P(\bar x/0)$, обозначим через $P_1(A)$ и $P_2(A)$ соответственно. Определим величину $W(F,\bar P)$ следующим образом:

$$W(F,\bar P)=\int\!\dots\!\int[P_2(\phi_1(\bar x))+\frac{1}{2}P_2(\phi_2(\bar x))]\,dP_1(\bar x).$$

В §I было показано, что потери для конкретной совокупности определяются выражением

$$Z(F(\bar{X}_1,\ldots,\bar{X}_\nu)) = \left[\varphi(\Pi) - \frac{(q+1)q}{2}\right]\tau.$$

ТЕОРЕМА 4.

$$M(Z(F,\bar{P})/q) = W(F,\bar{P})\tau(\nu - q)q.$$

ДОКАЗАТЕЛЬСТВО. Пусть набор $\{a_1,\ldots,a_\nu\}$ фиксирован. По определению,

$$Z(F(\bar{X}_1,\ldots,\bar{X}_\nu)) = \left[\sum_{i=1}^{\nu} \Delta(a_i) - \frac{q(q+1)}{2}\right]\tau .$$

Обозначим через σ_i' (соответственно σ_i'') количество объектов a_j таких, что $X_0(a_j)=0$ (соответственно $X_0(a_j)=1$) и j предшествует i в $F(\bar{X}_1,\ldots,\bar{X}_\nu)$. Тогда

$$Z(F(\bar{X}_1,\ldots,\bar{X}_\nu)) = \left[\sum_{i\in\{i|X_0(a_i)=1\}}(\sigma_i'+\sigma_i'') + q - \frac{q(q+1)}{2}\right]\tau .$$

Очевидно, что $\sum\limits_{i\in\{i|X_0(a_i)=1\}}\sigma_i''$ не зависит от $F(\bar{X}_1,\ldots,\bar{X}_\nu)$ и равна $\frac{(q-1)q}{2}$. Тогда

$$Z(F(\bar{X}_1,\ldots,\bar{X}_\nu)) = \Big(\sum_{i\in\{i|X_0(a_i)=1\}}\sigma_i'\Big)\tau$$

и

$$M(Z(F,\bar{P})/q) = M(Z(F(\bar{X}_1,\ldots,\bar{X}_\nu),\bar{P})/q) = \left[\sum_{i\in\{i|X_0(a_i)=1\}} M\sigma_i'\right]\tau . \quad (3)$$

Далее заметим, что $M\sigma_i' = \int\ldots\int M(\sigma_i'/\bar{X})\,dP_1(\bar{X})$, где $M(\sigma_i'/\bar{X})$ есть среднее число успехов из $\nu-q$ испытаний, когда вероятность успеха есть $P_2(\phi_1(\bar{x})) + \frac{1}{2}P_2(\phi_2(\bar{x}))$. Очевидно, что $M(\sigma_i'/\bar{X}) = (\nu-q)\times$ $\times[P_2(\phi_1(\bar{x})) + \frac{1}{2}P_2(\phi_2(\bar{x}))]$. Подставляя в (3) это выражение, находим

$$M(Z(F,\bar{P})/q) = (\nu-q)\,q\tau\,W(F).$$

Доказательство закончено.

Теорема 4 позволяет в дальнейшем рассматривать $W(F,\bar{P})$ как величину, линейно связанную с $M_\varphi(F,\bar{P})$, и строить для нее некоторые оценки. Важно, что независимо от величины q минимум $MZ(F,\bar{P})$ по F

достигается одновременно с минимумом $W(F,\bar F)$. В то же время выражение для $W(F,\bar F)$ просто вычисляется через $P(\bar x/0)$ и $P(\bar x/1)$, поэтому оно будет более удобным для оценивания.

Переходим к задаче упорядочивания совокупности объектов $\{a_1,\ldots,a_\nu\}$ на основе обучающей выборки $\{b_1,\ldots,b_m\}$. Ранее мы рассматривали задачу выбора порядка $\Pi(\bar x_1,\ldots,\bar x_\nu)$ в условиях, когда известны распределения $P(\bar x/0)$ и $P(\bar x/1)$. Естественно, что в реальных задачах мы не имеем такой информации, нам известна лишь обучающая матрица $\tilde X = \{x_{ij}\}$ $(i=1,\ldots,m;\ j=1,\ldots,n,0)$, где $x_{ij}=X_j(b_i)$, $\bar x_i = \bar X(b_i) = (X_1(b_i),\ldots,X_n(b_i))$, $\bar x_i^0 = \bar X^0(b_i) = (X_1(b_i),\ldots,X_n(b_i), X_0(b_i))$. Для объектов совокупности $\{b_1,\ldots,b_m\}$ мы знаем значения признаков и значения целевого признака X_0. Значения x_{ij} представляют выборочные значения для случайной многомерной величины с распределениями $P(\bar x/1)$ при $X_0(b_i) = 1$ и $P(\bar x/0)$ при $X_0(b_i) = 0$. Эта ситуация аналогична той, которая имеет место в статистической теории распознавания образов. При заданных распределениях статистическая теория принятия решений позволяет строить оптимальные решающие правила; основной результат здесь содержится в лемме Неймана-Пирсона. В задаче же распознавания распределения неизвестны, и нам необходимо строить решающее правило, основываясь на обучающей выборке, т.е. выборке, взятой из генеральной совокупности в соответствии с распределениями $P(\bar x/0)$, $P(\bar x/1)$ и априорными вероятностями $P(0)$ и $P(1)$.

Будем предполагать, что в задаче упорядочивания совокупности $\{a_1,\ldots,a_\nu\}$ фиксирован некоторый класс решающих правил $S = \{F\}$. Напомним, что F есть отображение:

$$(\bar x_1,\ldots,\bar x_\nu) \to (i_1,\ldots,i_\nu).$$

Рассматриваемая модель будет строиться по аналогии с моделями задач распознавания [I,5,6].

Заметим, что решающему правилу F можно поставить в соответствие порядок на элементах $\bar x \in X_1\cdot\ldots\cdot X_n$ следующим образом: если $\bar x = \bar X(a_i)$ и $\bar y = \bar X(a_j)$ и в совокупности $\Pi(\bar x_1,\ldots,\bar x_\nu) = (i_1,\ldots,i_\nu)$ i предшествует j, то полагаем $\bar x \succ \bar y$; если предшествование i и j определяется рандомизацией, то полагаем $\bar x \sim \bar y$.

ОПРЕДЕЛЕНИЕ 4. Отображение D, сопоставляющее каждой обучающей матрице $\tilde X$ некоторый элемент $F\in S$, будем называть алгоритмом упорядочивания (точнее было бы говорить алгоритм выбора упорядочивающего решающего правила).

Очевидно, что можно найти существенные аналогии между решающим правилом в распознавания и упорядочивающим решающим правилом, алгоритмом распознавания и алгоритмом упорядочивания [5].

Выше было установлено, что величина $W(F,\bar{P})$ линейно связана с $\varphi(F,\bar{P})$, следовательно, в классе S необходимо выбирать F такое, чтобы $W(F,\bar{P})$ была минимальной. Но мы не имеем возможности сделать это точно, поскольку нам неизвестна стратегия природы $\bar{P}$. Мы имеем лишь информацию о $\bar{P}$, содержащуюся в обучающей матрице $\tilde{X}$. Пусть n_1 и n_2 – соответственно количество объектов первого и второго образов в обучающей выборке, $n = n_1 + n_2$. Мы можем рассмотреть выборочные распределения $\hat{P}(\bar{x}/1)$ и $\hat{P}(x/0)$ и в качестве оценки для $W(F)$ величину

$$\hat{W}(F,\bar{P}) = \frac{1}{n_1}\cdot \sum_{i\in\{i\,|\,X_0(b_i)=1\}} \left[\frac{\sigma_i'}{n_2} + \frac{1}{2}\frac{\sigma_i''}{n_2}\right],$$

где σ_i' – количество объектов b_j таких, что $X_0(b_j) = 0$ и $\bar{x}(b_j) \succ \bar{x}(b_i)$, σ_i'' – количество объектов a_j таких, что $X_0(b_j) = 0$ и $\bar{x}(b_j) \approx \bar{x}(b_i)$.

Если мы будем рассматривать эту величину в качестве критерия выбора F из класса S, то перед нами встанет вопрос о том, в каком смысле эта оценка хорошо приближает величину $W(F,\bar{P})$. Здесь мы опять проводим параллель с задачей распознавания в том виде, в каком она рассматривается в работах В.Н.Вапника и А.Я.Червоненкиса [I,7]. Используя результаты указанных выше работ, мы исследуем некоторые формальные свойства предложенного подхода к упорядочиванию. Чтобы получить упорядочивающее решающее правило F, дающее значение критерия $W(F,\bar{P})$, близкое к $\inf_{F\in S} W(F,\bar{P})$, еще недостаточно выбрать стратегию, минимизирующую $\hat{W}(F,\bar{P})$.

Рассмотрим пример. Пусть $P(\bar{x}/1)$ и $P(\bar{x}/0)$ – невырожденные распределения, имеющие плотности $f(\bar{x}/1)$ и $f(\bar{x}/0)$. Пусть величина $W(F_{i\rho},\bar{P})$ достаточно мала и класс S таков, что в пространстве $X_1\cdot\ldots\cdot X_m$ могут быть заданы все упорядочения.

Рассмотрим следующий алгоритм упорядочения. Задается $\bar{x} \succ \bar{y}$, если существуют i и j такие, что $\bar{x} = \bar{x}(b_i)$, $\bar{y} = \bar{x}(b_j)$, $X_0(b_i) = 1$ и $X_0(b_j) = 0$ (где b_i, b_j – объекты обучающей выборки), а $\bar{x} \approx \bar{y}$ – во всех остальных случаях. Очевидно, что для любого объема выборки с вероятностью единица выполняется свойство

$$\hat{W}(F,\bar{P}) = 0, \qquad W(F,\bar{P}) = 1/2 ,$$

следовательно, выборочная характеристика $\hat{W}(F,\bar{P})$ без дополнительных ограничений не может служить оценкой величины $W(F,\bar{P})$.

Введем условия равномерного приближения $W(F,\bar{P})$ выборочной характеристикой $\hat{W}(F,\bar{P})$, опираясь на результаты В.Н.Вапника и А.Я.Червоненкиса [I, гл.X и 7] о равномерной сходимости величины эмпирического риска к истинному.

С каждым упорядочивающим решающим правилом F мы свяжем класс событий $O(F)$, который состоит из множеств вида $\{\bar{x} \in X_1 \cdot \ldots \cdot X_n \mid \bar{x} \succ_F \bar{x}_0\}$, где $\bar{x}_0 \in X_1 \cdot \ldots \cdot X_n$.

Определим систему событий $O(S) = \bigcup_{F \in S} O(F)$. Следуя [I, гл. X, §3,4], введем для данной системы событий понятия функции роста и емкости. Пусть $\Delta^{O(S)}(\bar{x}_1,\ldots,\bar{x}_n)$ есть число разбиений выборки событиями из $O(S)$. Функцию $m^{O(S)}(n) = \max \Delta^{O(S)}(\bar{x}_1,\ldots,\bar{x}_n)$ назовем функцией роста системы $O(S)$. Число

$$K(S) = \min \{n \mid m^{O(S)}(n) \neq 2^n\} - 1$$

назовем емкостью класса S.

Емкость класса $K(S)$ определяется единственным образом, так как, по основному свойству функции роста [I, гл. X, § 4], либо $m^{O(S)} = 2^n$, либо $m^{O(S)} \le \sum_{i=0}^{K(S)} C_n^i$.

В работе [7] получена оценка вероятности равномерного отклонения величины

$$I(\alpha) = M(Q(\bar{x},\alpha)) = \int Q(\bar{x},\alpha)\, dP(\bar{x})$$

от величины эмпирического функционала

$$I_{\text{эмп}}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} Q(\bar{x}_i,\alpha).$$

Если $0 \le Q(\bar{x},\alpha) \le K$ (K не зависит от $\bar{x}$ и α), то

$$P\left\{\sup_{\alpha \in S} |I(\alpha) - I_{\text{эмп}}(\alpha)| > \varepsilon\right\} < 2m^S(2n) \exp\left(\frac{-\varepsilon^2 n}{4K^2}\right). \qquad (4)$$

Здесь $m^S(2n)$ – функция роста для системы событий вида $\{Q(\bar{x},\alpha) \le C\}$. С помощью этого результата доказывается

ТЕОРЕМА 5. Имеет место оценка вероятности равномерного отклонения $W(F,\bar{P})$ от $\hat{W}(F,\bar{P})$ вида

$$P_{\bar{P}}(\sup_F |W(F,\bar{P}) - \hat{W}(F,\bar{P})| > 2\varepsilon) \le 2m^{O(S)}(2n_1)\exp\left(\frac{-\varepsilon^2 n_1}{4}\right) +$$

$$+ 2m^{O(S)}(2n_2)\exp\left(\frac{-\varepsilon^2 n_2}{4}\right) -$$

$$- 4m^{O(S)}(2n_1)m^{O(S)}(2n_2)\exp\left(\frac{-\varepsilon^2(n_1+n_2)}{4}\right) \qquad (5)$$

для каждой стратегии природы $\bar{P} = (P(\bar{x}/1), P(\bar{x}/0))$.

ДОКАЗАТЕЛЬСТВО. Пусть

$$Q(\bar{x},F) = P_2(\bar{x},F) = P_2(\phi_1(\bar{x})) + \frac{1}{2}P_2(\phi_2(\bar{x})) .$$

Мы можем считать, что sup в (4) берется не по параметру α, а по элементам F из некоторого множества S , т.е. множество S может быть задано и непараметрически. Важно, чтобы каждому F соответствовало свое значение функционала $I(F)$. Доказательство утверждения (4), приводимое в [7], не претерпит никаких изменений, если рассматривать равномерное отклонение в таком виде.

Пусть $I_1 = \{i | X_0(a_i) = 1\}$. Тогда (4) дает

$$P_1 \left\{\sup \left| \int\ldots\int P_2(\bar{x},F)dP_1(\bar{x}) - \right.\right.$$

$$\left.\left. - \frac{1}{n_1}\sum_{i\in I_1} P_2(x_i,F)\right| > \varepsilon_1\right\} \le 2m^{O(S)}(2n_1)\exp\left[\frac{-\varepsilon_1^2 n_1}{4}\right] \qquad (6)$$

для каждой стратегии природы $\bar{P} = (P(\bar{x}/1), P(\bar{x}/0))$.

Положив в (4) $Q(\bar{x}, \bar{x}_i, F) = \begin{cases} 1 & \text{при } \bar{x} \succ \bar{x}_i, \\ 1/2 & \text{при } \bar{x} = \bar{x}_i, \\ 0 & \text{при } \bar{x} \prec \bar{x}_i, \end{cases}$

мы можем показать, что

$$P_2\left\{\sup_{\bar{x}_i,F}\left|P_2(\bar{x}_i,F)-\frac{1}{n_2}(\sigma_i'+\frac{1}{2}\sigma_i'')\right|>\epsilon_2\right\}\le$$

$$\le 2m^{O(S)}(2n_2)\exp\left(\frac{-\epsilon_2^2 n_2}{4}\right).$$

Следовательно, для каждого набора I_1 выполняется

$$P_2\left\{\sup_F\left|\frac{1}{n_1}\sum_{i\in I_1}P_2(\bar{x}_i,F)-\right.\right.$$

$$\left.\left.-\frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i'+\sigma_i'')\right|>\epsilon_2\right\}\le 2m^{O(S)}(2n_2)\exp\left(\frac{-\epsilon_2^2 n_2}{4}\right). \qquad (7)$$

Если

$$\left|\int P_2(\bar{x},F)dP_1(\bar{x})-\frac{1}{n_1}\sum_{i\in I_1}P_2(\bar{x}_i,F)\right|\le\epsilon$$

и

$$\left|\frac{1}{n_1}\sum_{i\in I_1}P_2(\bar{x}_i\ F)-\frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i'+\frac{1}{2}\sigma_i'')\right|\le\epsilon,$$

то

$$\left|\int P_2(\bar{x},F)dP_1(\bar{x})-\frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i'+\frac{1}{2}\sigma_i'')\right|\le 2\epsilon.$$

Поскольку события в неравенствах (6) и (7) независимы, то

$$P\left\{\sup_F\left|\int P_2(\bar{x})dP_1(\bar{x})-\frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i'+\frac{1}{2}\sigma_i'')\right|\le 2\epsilon\right\}\ge$$

$$\ge P\left[\left\{\sup_F\left|\int P_2(\bar{x},F)dP_1(\bar{x})-\frac{1}{n_1}\sum_{i\in I_1}P_2(\bar{x}_i,F)\right|\le\epsilon\right\}\cap\right.$$

$$\left.\cap\left\{\sup_{(\bar{x}_i,F)}\left|\frac{1}{n_1}\sum_{i\in I_1}P_2(x_i\ F)-\frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i'+\frac{1}{2}\sigma_i'')\right|\le\epsilon\right\}\right]=$$

$$=P\left\{\sup_F\left|\int P_2(\bar{x},F)dP_1(\bar{x})-\frac{1}{n_1}\sum_{i\in I_1}P_2(\bar{x}_i,F)\right|\le\epsilon\right\}\times$$

$$\times P\left\{\sup_{(\bar{x}_i,F)}\left|\frac{1}{n_1}\sum_{i\in I_1}P_2(\bar{x}_i,F)-\frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i'+\frac{1}{2}\sigma_i'')\right|\le\epsilon\right\}>$$

$$>(1 - 2m^{O(S)}(2n_1)\exp\left(\frac{-\varepsilon^2 n_1}{4}\right))(1 - 2m^{O(S)}(2n_2)\exp\left(\frac{-\varepsilon^2 n_2}{4}\right)).$$

Отсюда следует:

$$P\left\{\sup_F\left|\int P_2(\bar{x},F)dP_1(\bar{x}) - \frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i' + \frac{1}{2}\sigma_i'')\right| > 2\varepsilon\right\} \le$$

$$\le 2m^{O(S)}(2n_1)\exp\left(\frac{-\varepsilon^2 n_1}{4}\right) + 2m^{O(S)}(2n_2)\exp\left(\frac{-\varepsilon^2 n_1}{4}\right) -$$

$$- 4m^{O(S)}(2n_1)\,m^{O(S)}(2n_2)\exp\left(\frac{-\varepsilon^2(n_1+n_2)}{4}\right).$$

Доказательство закончено.

Таким образом, мы получили оценку для вероятности равномерного отклонения величины $W(F,\bar{P})$ от $\hat{W}(F,P)$. Это дает возможность обоснованно применять алгоритмы упорядочения, для которых емкостная характеристика $E(S)$ соответствующего класса событий конечна. В [I, гл.X, §4] показана при $E(S)\ge 1$ и $n>E(S)$ справедливость неравенства

$$m^{O(S)}(n) \le 1.5\,\frac{n^{E(S)-1}}{E(S)!}.$$

Подставляя это неравенство в (5), имеем

$$P\{\sup_F\left|\int P_2(\bar{x},F)dP_1(\bar{x}) - \frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i' + \frac{1}{2}\sigma_i'')\right| > 2\varepsilon\} \le$$

$$\le 3\,\frac{n_1^{E(S)-1}}{E(S)!}\exp\left(-\frac{\varepsilon^2 n_1}{4}\right) + 3\,\frac{n_2^{E(S)-1}}{E(S)!}\exp\left(-\frac{\varepsilon^2 n_2}{4}\right) -$$

$$- 6\,\frac{(n_1 n_2)^{E(S)-1}}{(E(S)!)^2}\exp\left(-\frac{\varepsilon^2(n_1+n_2)}{4}\right).$$

Очевидно, что при n_1 и n_2 , стремящихся к бесконечности, правая часть стремится к 0, и теперь легко доказывается

ТЕОРЕМА 6.

$$\lim_{n_1,n_2\to\infty} P\{\sup_F |\int F_2(\bar{x},F)dF_1(\bar{x}) - \frac{1}{n_1 n_2}\sum_{i\in I_1}(\sigma_i' + \frac{1}{2}\sigma_i'')| > \varepsilon\} = 0$$

для каждого ε.

Пусть F_0 таково, что $W(F_0,\bar{F}) \le W(F,\bar{F})$ для каждого F из S, и $\hat{F}_n$ таково, что $\hat{W}(\hat{F}_n,F) \le \hat{W}(F,F)$ для каждого F из S.

Тогда непосредственно из теоремы 6 получаем

СЛЕДСТВИЕ.

$$P(|W(\hat{F}_n) - W(F_0)| > 4\varepsilon) \le$$

$$\le 2m^{O(S)}(2n_1)\exp\left(-\frac{\varepsilon^2 n_1}{4}\right) + 2m^{O(S)}(2n_2)\exp\left(-\frac{\varepsilon^2 n_2}{4}\right) -$$

$$- 4m^{O(S)}(2n_1)m^{O(S)}(2n_2)\exp\left(-\frac{\varepsilon^2(n_1+n_2)}{4}\right)$$

и

$$\lim_{n_1,n_2\to\infty} P(|W(\hat{F}_n) - W(F_0)| > \varepsilon) = 0$$

для каждого ε.

Это следствие показывает, что упорядочивающее решающее правило, которое мы выбираем по обучающей выборке $\{b_1,\ldots,b_n\}$, приближается к искомому.

Алгоритмы, реализующие описанные методы упорядочения, основаны на логических решающих функциях. Каждое логическое решающее правило определяет набор ветвей $\{B_1,\ldots,B_k\}$. Каждой ветви соответствует оценка функции правдоподобия $\hat{g}(B_i) = \frac{n_i+1}{m_i+1}$ (n_i – число объектов обучающей выборки, попавших в B_i из первого образа, m_i – соответствующее число объектов второго образа). Каждый объект a из совокупности $\{a_1,\ldots,a_y\}$, подлежащей упорядочению, попадает в некоторую ветвь $B_k = B(a_i)$.

Определим $g(a_i) = g(B_k) = g(B(a_i))$. Упорядочение задается следующим образом:

$$a_i \prec a_j,\ \text{если}\ g(a_i) < g(a_j),$$
$$a_i \cong a_j,\ \text{если}\ g(a_i) = g(a_j),$$
$$a_i \succ a_j,\ \text{если}\ g(a_i) > g(a_j).$$

Множество деревьев фиксированной длины определяет класс упорядочивающих функций S .

Решена задача по упорядочению четырехсот геологических участков по обучающей выборке, содержащей около трехсот объектов. С алгоритмами и программами можно ознакомиться по отчетам Института математики СО АН СССР.

Литература

1. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. Теория распознавания образов. М., "Наука", 1974.

2. ЛЕМАН Э. Проверка статистических гипотез. М.,"Наука",1964.

3. КЕНДАЛЛ М. Статистические выводы и связи. М.,"Наука",1973.

4. БЛЕКУЭЛ Д., ГИРШИК Н.А. Теория игр и статистических решений. М.,ИЛ, 1958.

5. ЛБОВ Г.С., МАНОХИН А.Н. Распознавание образов для разнотипных признаков в условиях малой выборки. -В кн.: Статистические проблемы управления. Вып.14. Вильнюс, 1976, с.49-57.

6. МАНОХИН А.Н. Методы распознавания образов, основанные на логических решающих функциях. -В кн.: Эмпирическое предсказание и распознавание образов. (Вычислительные системы. Вып. 67.) Новосибирск, 1976, с. 42-53.

7. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. Теория равномерной сходимости частот к их вероятностям и задача поиска оптимальных решений по эмпирическим данным. -"Автоматика и телемеханика", 1971, № 2, с. 42-54.

Поступила в ред.-изд.отд.
17 января 1978 года